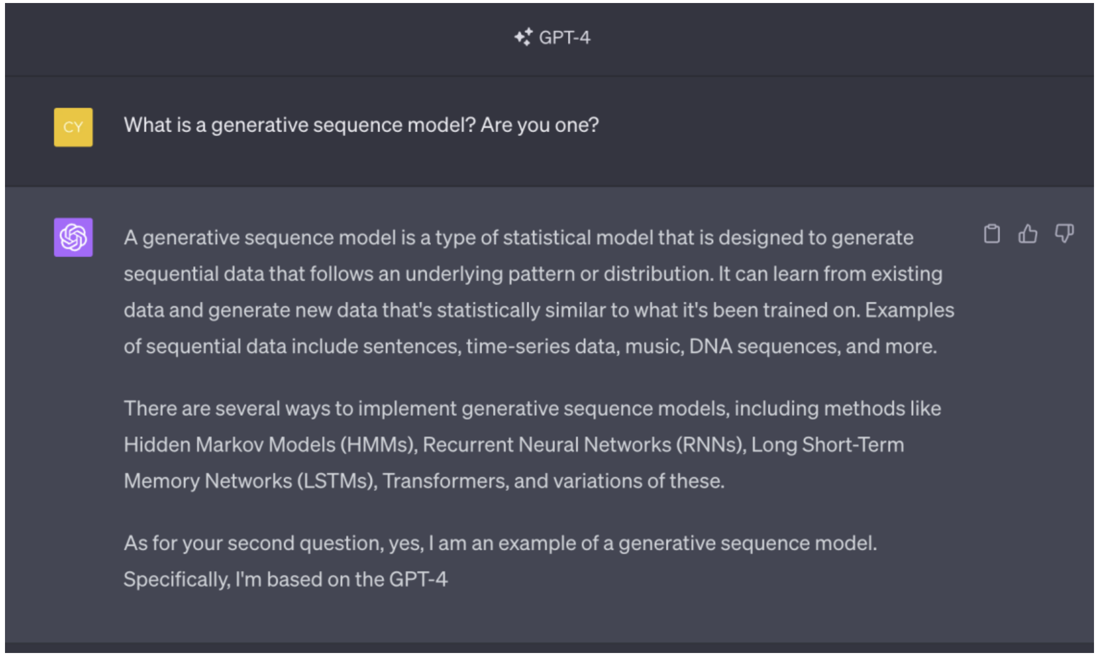
Learning Hidden Markov Models with conditional sampling

Akshay Krishnamurthy

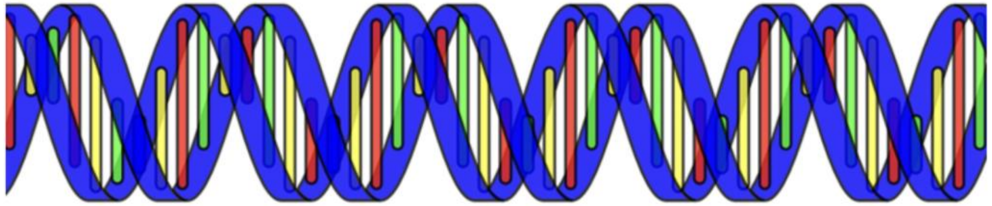
Microsoft Research, NYC

Joint work with Sham Kakade, Gaurav Mahajan, Cyril Zhang

Statistical modeling of sequences

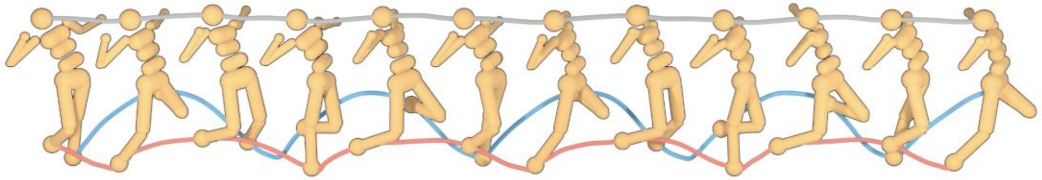


Large language models



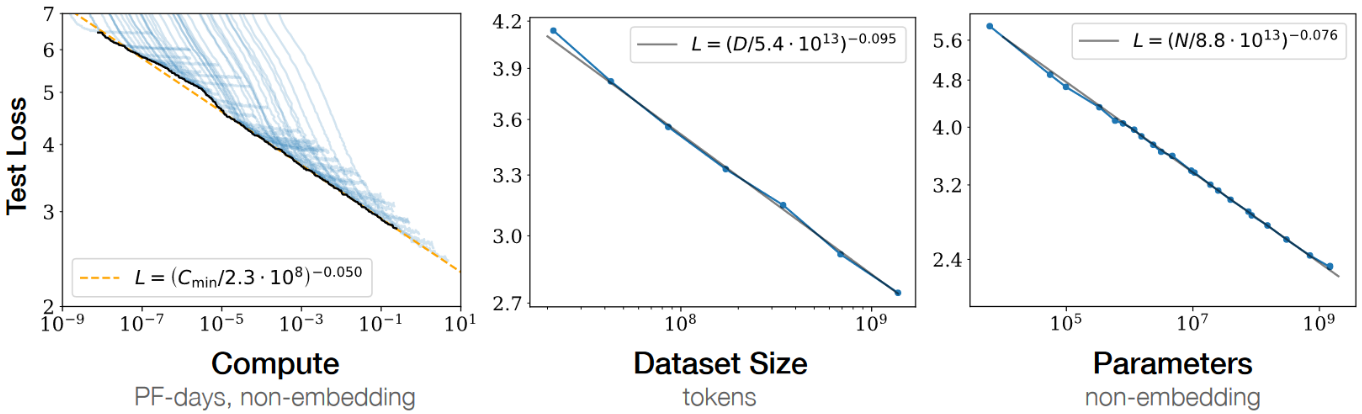
Biology

Speech/Audio



Robotics and decision making

Computational cost?



User: What is 8493 + 2357?
GPT-3.5: 10850 ✓
GPT-4: 10850 ✓

User: What is 84935834 + 23572898?
GPT-3.5: 108008732 ✗
GPT-4: 108508732 ✓

User: What is 9991999919909993 + 6109199190990097?
GPT-3.5: 16111199190810090 ✗
GPT-4: 16101199100890090 ✗
Answer: 16101199110900090

“data requirements growing very slowly as $D \sim C^{0.27}$ with compute”

Sequence modeling is computationally hard particularly for algorithmic/combinatorial problems

Can we circumvent these barriers?
What are the algorithmic principles for doing so?

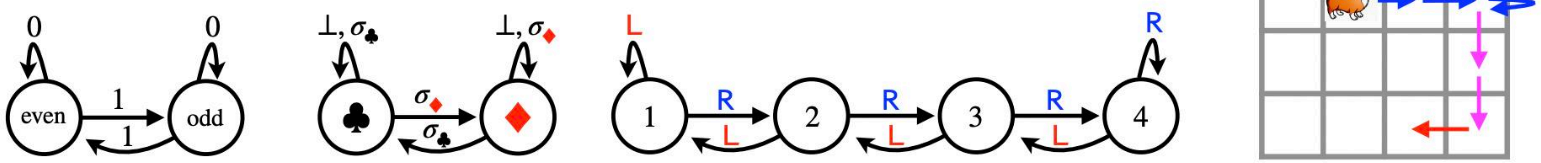
Kaplan, McCandlish et al. Scaling laws for neural language models. 2020
Liu, Ash, Goel, K, Zhang. Exposing attention glitches with flip-flop language modeling. 2023

Computational barriers in sequence modeling

Learning parity (with noise)

Learning deterministic finite automata

parity counter | memory unit | 1D gridworld



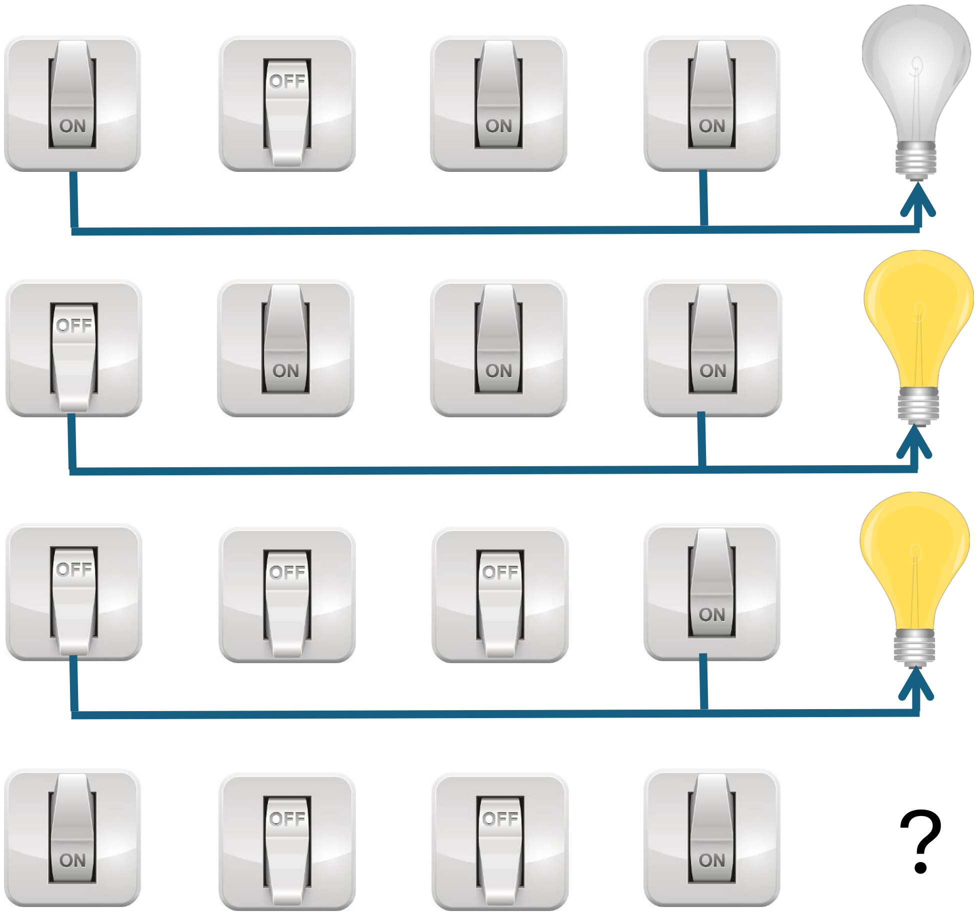
$Q = \{\text{even}, \text{odd}\}$
$\Sigma = \{0, 1\}$

$Q = \{♣, ♦\}$
$\Sigma = \{\sigma_♣, \sigma_♦, \bot\}$

$Q = \{1, 2, 3, 4\}$
$\Sigma = \{\text{L}, \text{R}\}$

$Q = \{1..3\} \times \{1..4\}$
$\Sigma = \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$

Vignette #1: learning parity (with noise)



Planted subset $S \subset [n]$ of size k

Samples $(x, y) \in \{-1,1\}^n \times \{-1,1\}$

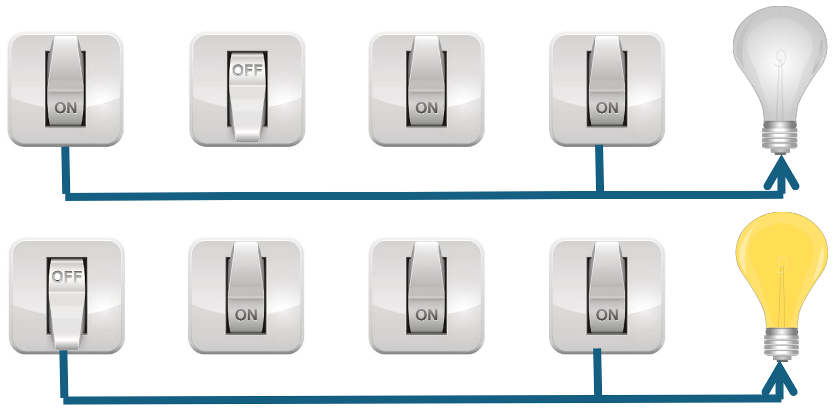
- $x \sim \mathrm{Unif}(\{-1,1\}^n)$
- $y = \prod_{i \in S} x_i$

Find classifier $h\colon \{-1,1\}^n \to \{-1,1\}$ minimizing

$$\Pr[\, h(x) \neq y \,]$$

Noisy version: $y = \prod_{i \in S} x_i$ with prob 3/4

Parity: theoretical results



Statistical complexity: $\Theta(k \log n)$ for k-sparse parity

$\binom{n}{k}$ hypothesis, all wrong ones have error rate ½

Parity functions are orthogonal: $\mathbb{E}[\,\chi_S(x)\chi_T(x)] = 1\{\,T = S\,\}$

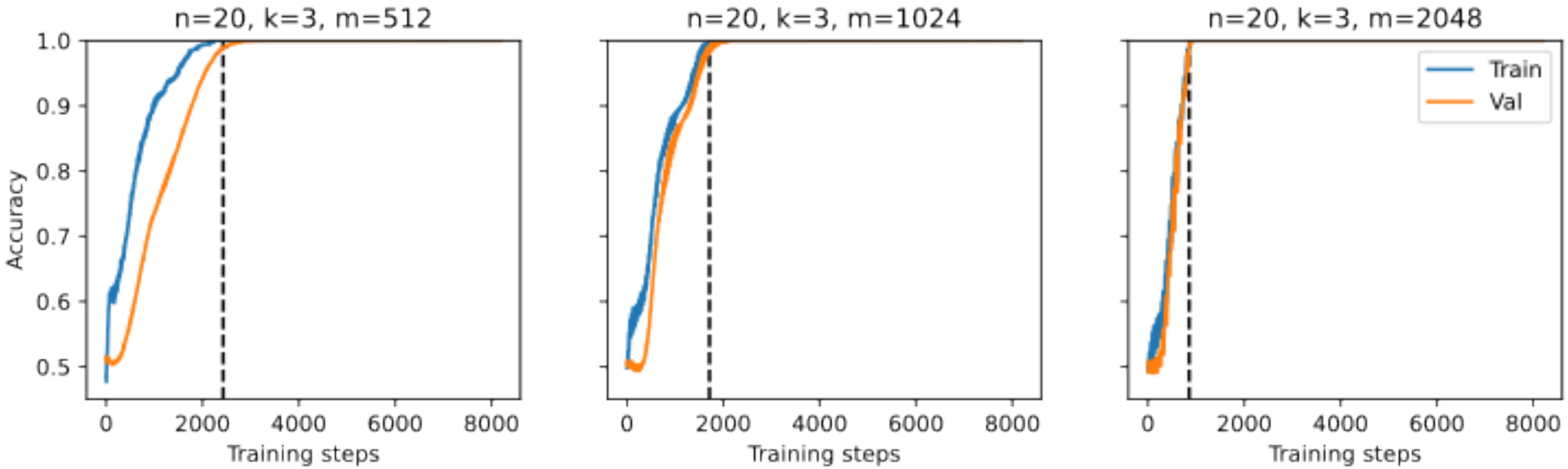
Computational complexity:

- Noiseless parity: poly(n) time via Gaussian elimination
- Noisy case: try all $\binom{n}{k} \sim n^k$ hypotheses
- $\Omega(n^k)$ time via statistical queries
- Noisy parity: conjectured $n^{\Omega(k)}$ for all algorithms

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Kearns. Efficient noise tolerant learning from statistical queries. 1998

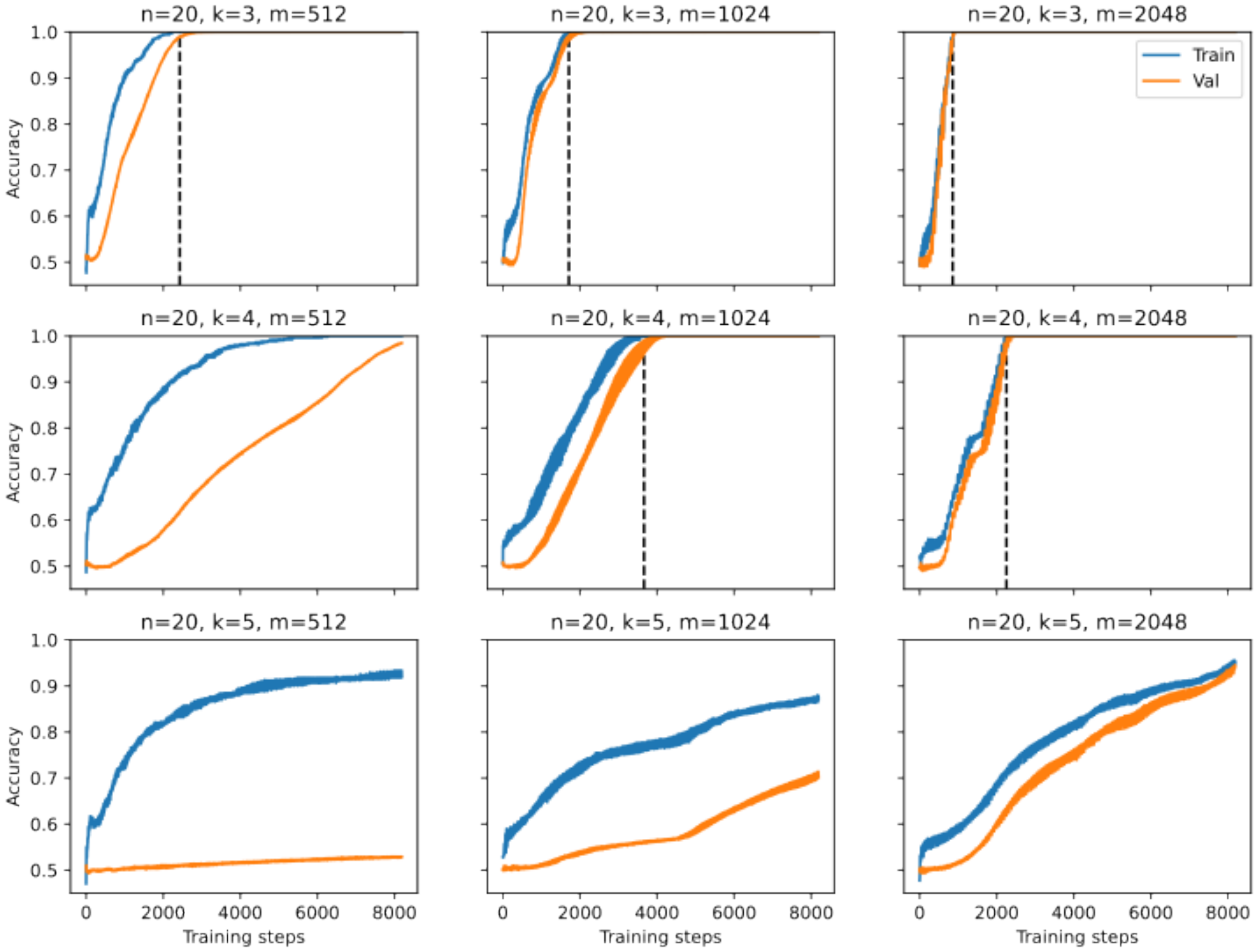
Experiments with parity



Accuracy curves shift left with more samples => comp-stat tradeoff

Barak. Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit. 2023.
Abbe. Provably advantage of curriculum learning on parity targets with mixed inputs. 2023.

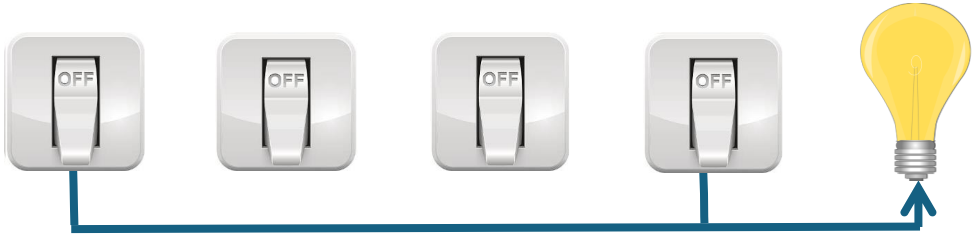
Experiments with parity



Problem gets much harder as k increases

Barak et al. Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit. 2023.
Abbe et al. Provably advantage of curriculum learning on parity targets with mixed inputs. 2023.

Interactive learning for sparse parity



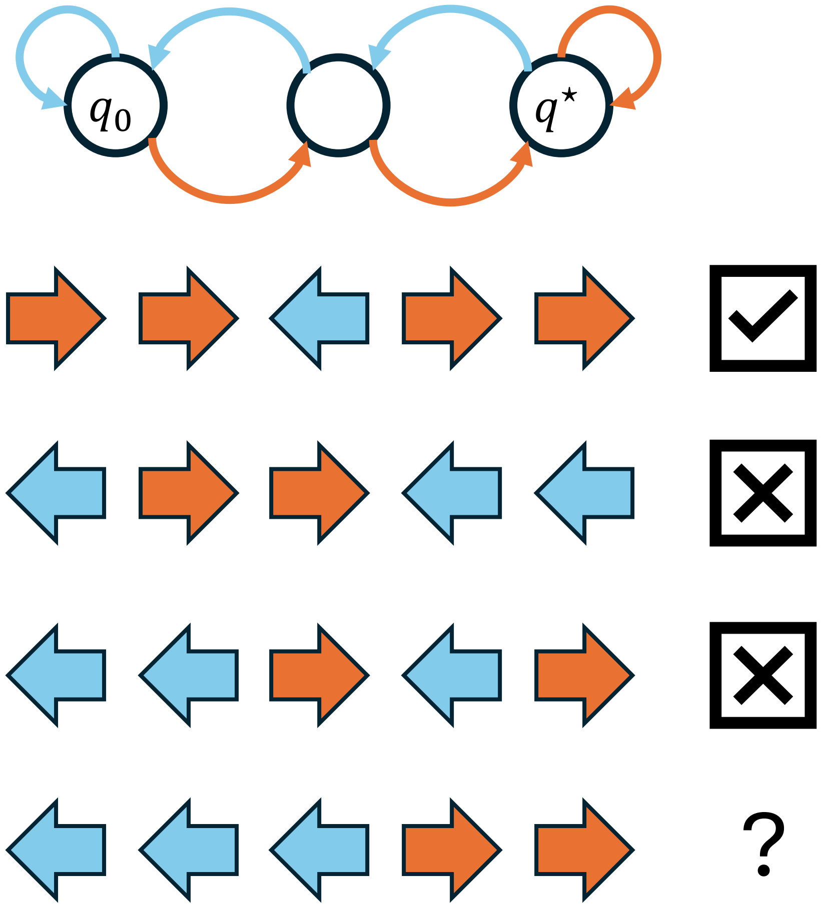
Pick arbitrary x, get label y
For each $i \in [n]$: query $x^i = (x_1, \ldots, 1 - x_i, \ldots, x_n)$ get label y^i

$O(n)$ queries/samples and $O(n)$ time

- Works with noise
- Works for any sparsity

"interactive learning" bypasses computational hardness for parity

Vignette #2: Deterministic finite automata



Deterministic Finite Automata:

- Finite state space Q, start state q_0, goal $\mathrm{q}^\star$
- Finite input alphabet Σ
- Transition operator $\mathrm{T}: \mathrm{Q} \times \Sigma \to \mathrm{Q}$

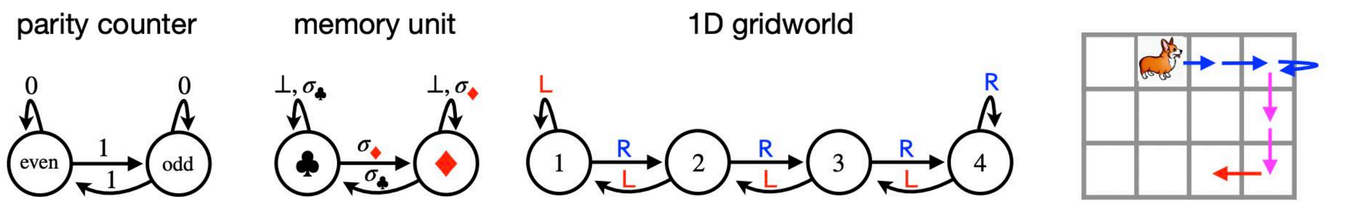
Learning problem: unknown DFA with n states
Samples $(\mathrm{x},\mathrm{y}) \in \Sigma^{\mathrm{T}} \times \{0,1\}$

- $\mathrm{x} \sim \mathrm{Unif}(\Sigma^{\mathrm{T}})$
- $\mathrm{y} = 1\{\mathrm{q}_{\mathrm{T}}(\mathrm{x}) = \mathrm{q}^\star\}$, where $\mathrm{q}_{\mathrm{t}}(\mathrm{x}) = \mathrm{T}(\mathrm{q}_{\mathrm{t}-1}(\mathrm{x}), \mathrm{x}_{\mathrm{t}})$

Find classifier $\mathrm{h}: \Sigma^{\mathrm{T}} \to \{0,1\}$ minimizing

$$\Pr[\,\mathrm{h}(\mathrm{x}) \neq \mathrm{y}\,]$$

DFAs: theory



$Q = \{\text{even}, \text{odd}\}$, $\Sigma = \{0, 1\}$

$Q = \{♣, ♦\}$, $\Sigma = \{\sigma_♣, \sigma_♦, \bot\}$

$Q = \{1, 2, 3, 4\}$, $\Sigma = \{L, R\}$

$Q = \{1..3\} \times \{1..4\}$, $\Sigma = \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$

Statistical complexity:

$\sim Q^{Q\times\Sigma+2}$ DFAs with states Q and inputs Σ

$\Rightarrow$ learn ϵ approximation with $O(Q\Sigma \log(Q)/\epsilon)$ samples

Computational complexity:

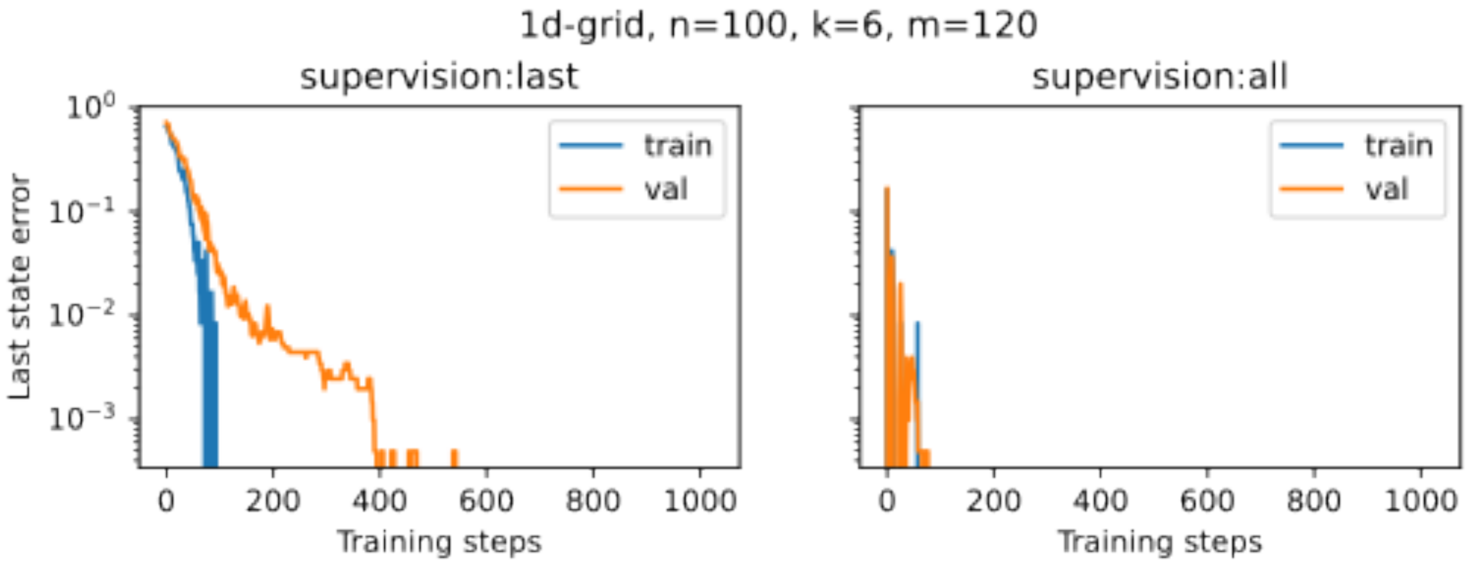
- Not efficiently learnable under discrete cube root hypothesis
- Efficiently learnable with membership and equivalence queries!

“interactive learning” bypasses computational hardness for DFAs

Kearns and Valiant. Cryptographic limitations on learning Boolean formulae and finite automata. 1994
Angluin. Learning regular sets from queries and counterexamples. 1987

Experiments with DFAs

Natural to study DFAs in the context of neural algorithmic reasoning



Hard to train recurrent models: vanishing gradients lead to slow convergence
Mitigated with other architectures, but new issues crop up: shortcuts

Liu, Ash, Goel, K, Zhang. Transformers learn shortcuts to automata. 2022.

Statistical sequence modeling

Sequence modeling is computationally hard particularly for algorithmic/combinatorial problems

Can we circumvent these barriers?
What are the algorithmic principles for doing so?

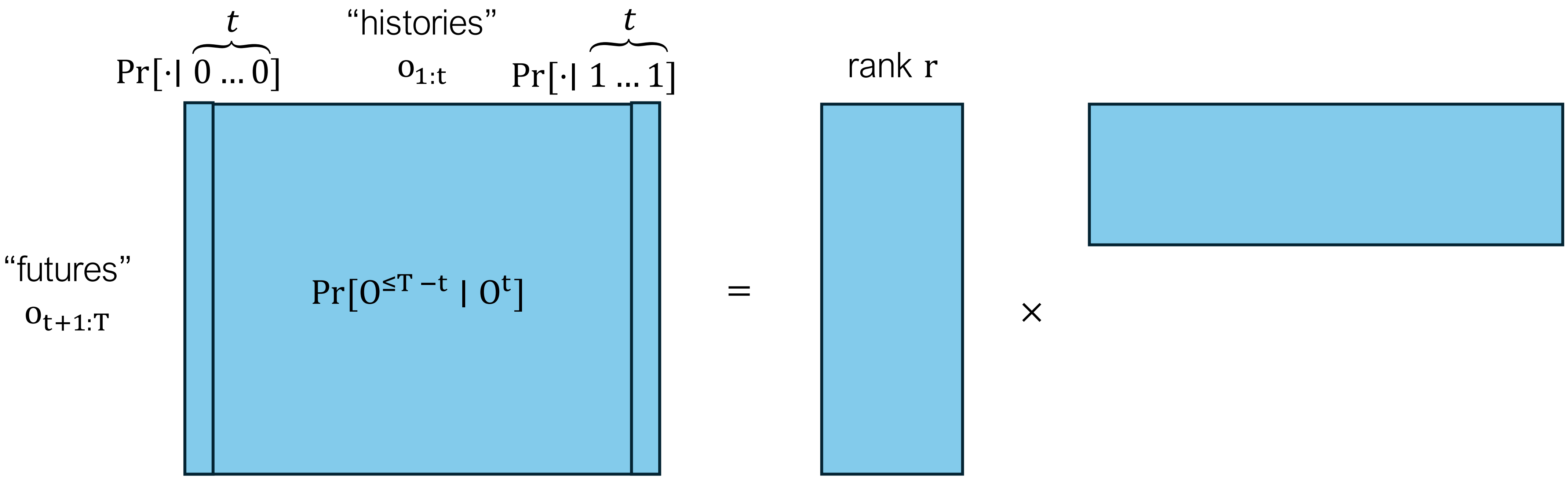
For parity, DFAs:

- ✓ Yes via interactive learning!
- ✗ Problem structure is very specific/discrete
- ✗ Highly specialized algorithm design

Is there a more general statistical model for studying these questions?

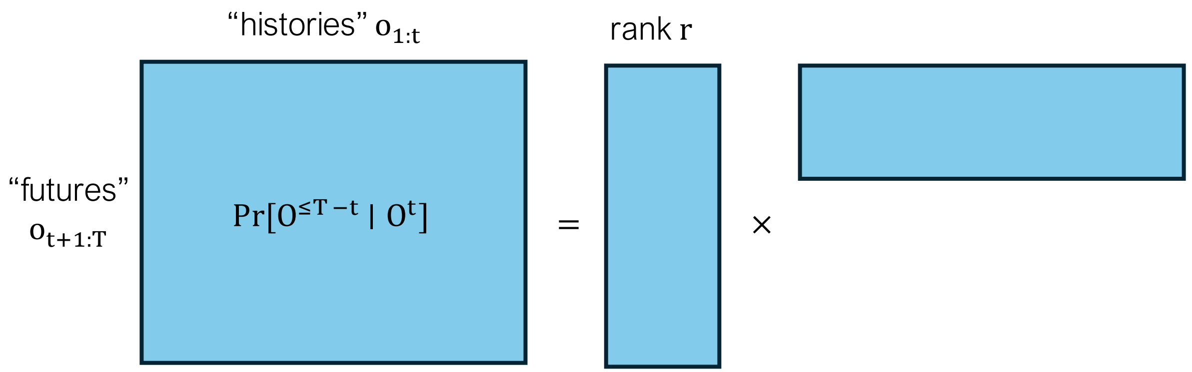
Low rank distributions

Distribution $\Pr[\cdot]$ over sequences O^T for finite observation space O



For each t: $\Pr[\text{Futures}_t \mid \text{Histories}_t]$ matrix has rank at most r.
Note: matrices (and factorization) are exponentially large!

Low rank distributions



Special cases:
- Parity (with noise) has rank 4
- DFAs have rank at most Q, the number of states
- Subsumes Hidden Markov models: rank at most S, the number of hidden states

Learning goal: Efficiently output $\widehat{\Pr}[\cdot]$ that ϵ approximates $\Pr[\cdot]$ in total variation distance

$$\frac{1}{2} \cdot \sum_{o_1, \ldots o_T} \left| \Pr[o_{1:T}] - \widehat{\Pr}[o_{1:T}] \right| \leq \epsilon$$

Efficiently: w.p. $1 - \delta$ in time $\mathrm{poly}(r, T, O, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$.

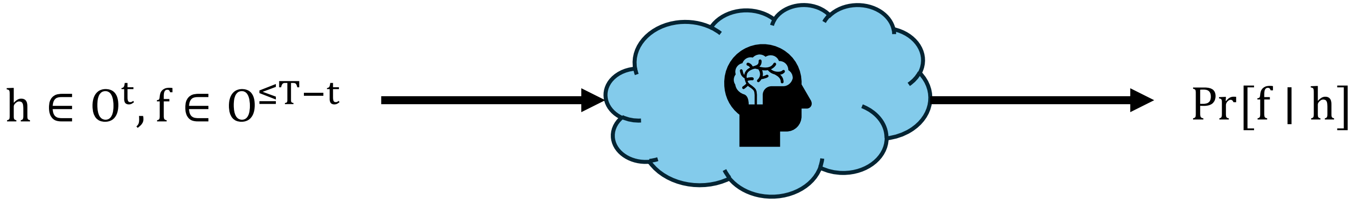
Learning low rank distributions from samples $o_{1:T} \sim \Pr[\cdot]$ is computationally hard!

Interactive learning models

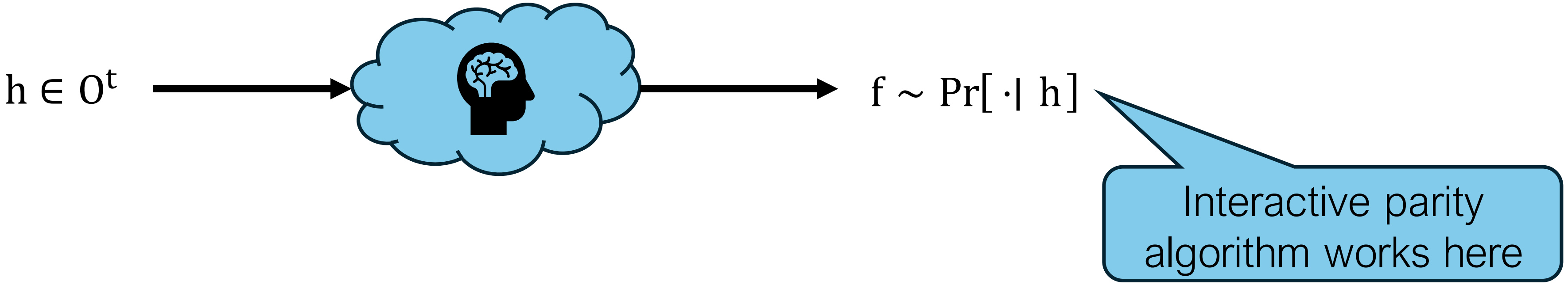
Non-interactive model: iid samples $o_{1:T} \sim \Pr[\cdot]$

- Computationally hard: Parity, DFA, HMM all special cases

Interactive model #1: Exact conditional oracle reveals $\Pr[f \mid h]$ on inputs (h, f).



Interactive model #2: Conditional sampling oracle samples $f \sim \Pr[\cdot \mid h]$ on input h.



Chakraborty et al. On the power of conditional samples in distribution testing. 2013

Our results

Theorem 1: poly$(\mathrm{r}, \mathrm{T}, \mathrm{O}, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$ time algorithm for rank r distributions using iid samples and exact conditional probability oracle

Generalizes Angluin's L* algorithm for learning DFAs.

Theorem 2: poly$(\mathrm{r}, \mathrm{T}, \mathrm{O}, \frac{1}{\Delta}, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$ time algorithm for rank r distributions using iid samples and conditional sampling oracle.

Δ is the fidelity of the distribution (defined later). Open problem!
Captures parity, prior results for HMMs, but not DFAs

Proof overview

- Structural properties: compact representation via observable operators
  - Generalizing Angluin's L* algorithm with exact conditional probabilities
- Estimating operators
- Basis finding
- Error propagation

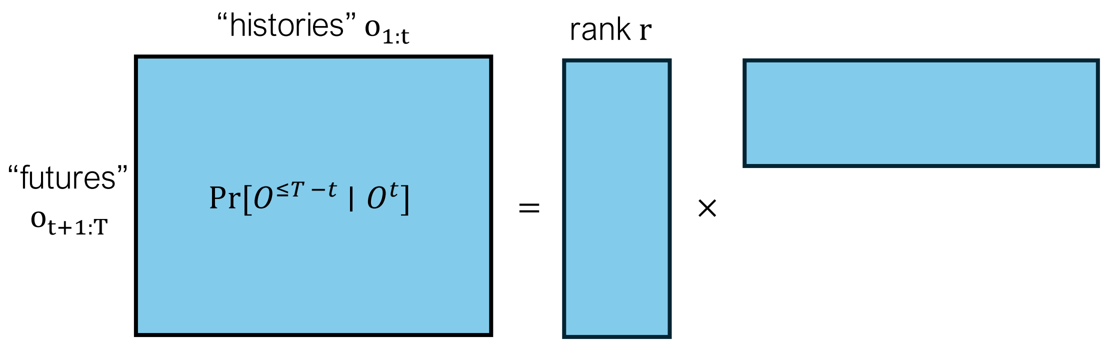
Proof overview

- Structural properties: compact representation via observable operators
    - Generalizing Angluin's L* algorithm with exact conditional probabilities
- Estimating operators
- Basis finding
- Error propagation

Compact Representation

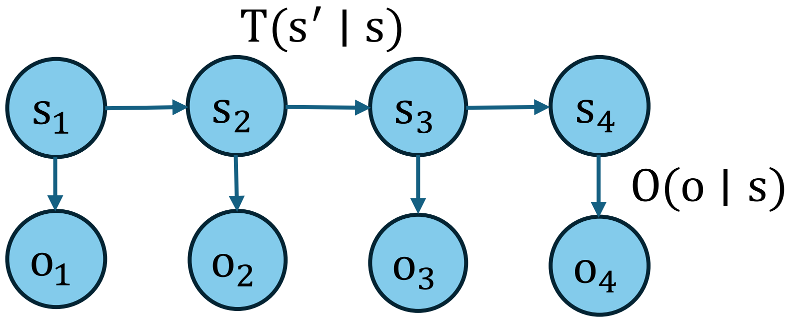
Basic question: $\Pr[\cdot]$ has support O^T. Can we efficiently write down estimate $\widehat{\Pr}[\cdot]$?

Idea #1: Use low rank assumption



Factors are still exponentially large!

Idea #2: use explicit parametrization



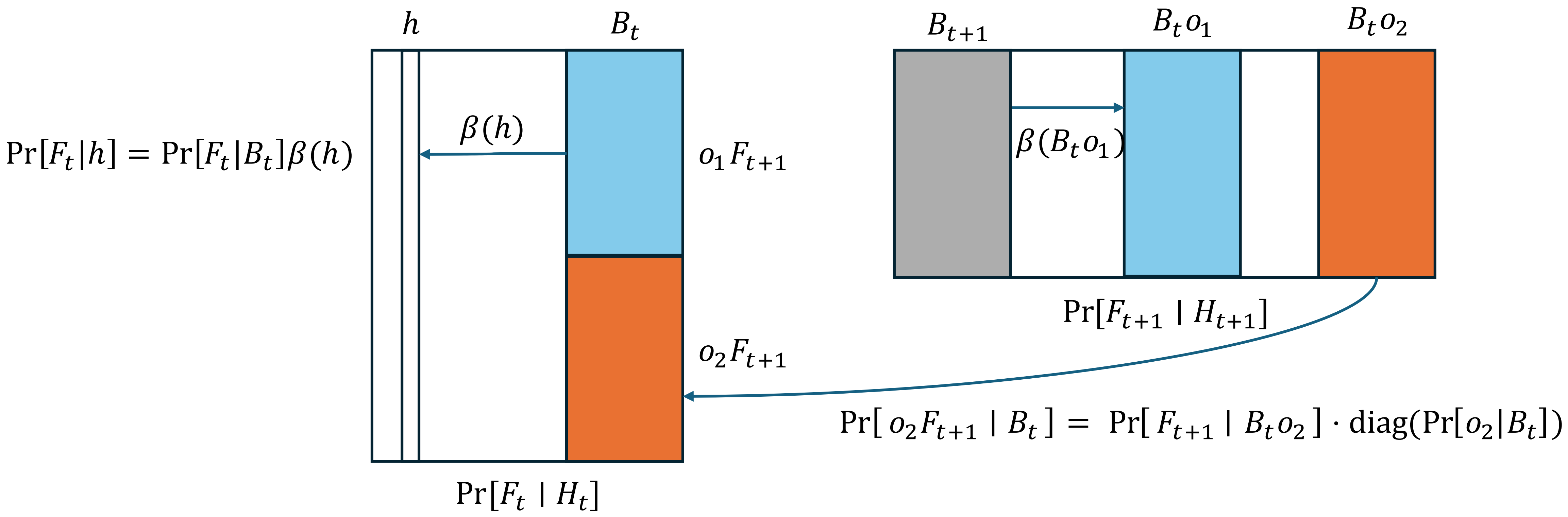
Only plausible for HMMs/DFAs/etc.

Proposition: For any rank r distribution, there exist matrices $\{A_{o,t}\}_{o\in O, t\in\{0,\ldots,T-1\}}$ of size at most $r \times r$ such that

$$\forall x_1, \ldots x_T \in O^T\colon \Pr[x_1, \ldots x_T] = A_{x_T,T-1}\, A_{x_{T-1},T-2} \ldots A_{x_1,0}$$

Jaeger. Observable operator models for discrete stochastic time series. 2000

Compact representation 1: basis histories



Each matrix has a "basis" of r histories B_t, which can linearly represent all columns

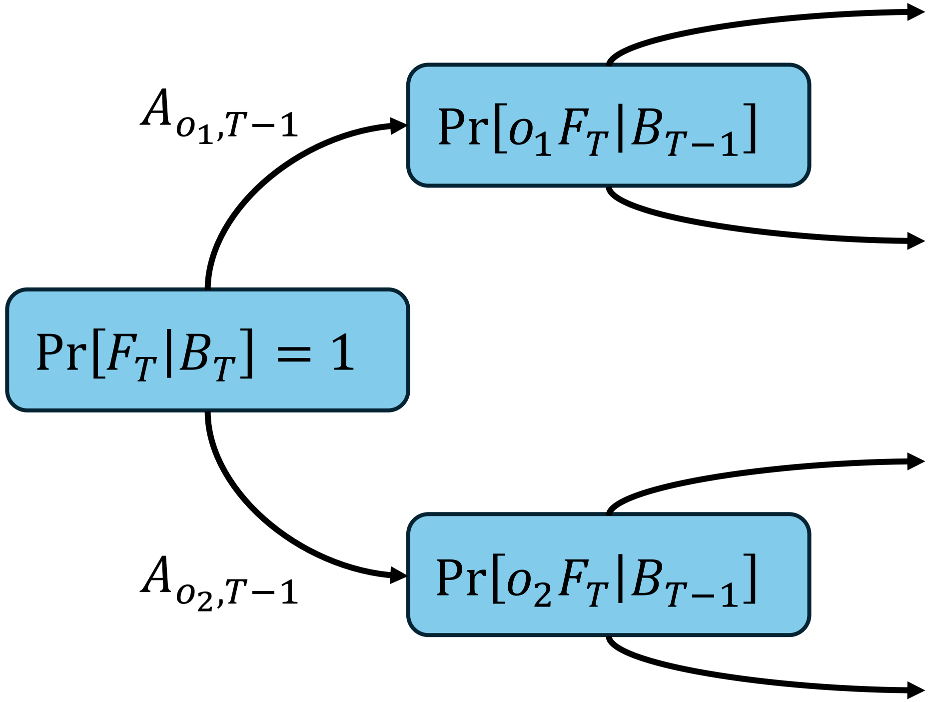
Basis submatrices appear in matrix at the next time! (up to rescaling)

Can write basis at time t as a function of basis at time $t+1$

Compact representation 2: operators

$$\Pr[\, o\, F_{t+1} \mid B_t \,] = \Pr[\, F_{t+1} \mid B_t\; o \,] \cdot \mathrm{diag}(\Pr[o|B_t]) = \Pr[\, F_{t+1} \mid B_{t+1} \,] \cdot \underbrace{\beta(B_t o) \cdot \mathrm{diag}(\Pr[o|B_t])}_{A_{o,t}}$$

If $A_{o,t}$s are known, we can inductively compute basis matrices!
Base case: $F_T = \emptyset$ so $\Pr[F_T|B_T] = 1$



$\Pr[o_1 \ldots o_1|B_0]$

$\Pr[o_1 \ldots o_2|B_0]$

Full sequence probabilities!

$\Pr[o_2 \ldots o_1|B_0]$

$\Pr[o_2 \ldots o_2|B_0]$

Compact representation 3: operators

Proposition: For any rank r distribution, there exist matrices $\{A_{o,t}\}_{o \in O, t \in \{0,\ldots,T-1\}}$ of size at most $r \times r$ such that

$$\forall x_1, \ldots x_T \in O^T: \Pr[x_1, \ldots x_T] = A_{x_T,T-1} A_{x_{T-1},T-2} \ldots A_{x_1,0}$$

$A_{o,t}$ is the solution to the equation $\Pr[F_{t+1}|B_{t+1}]A_{o,t} = \Pr[oF_{t+1}|B_t]$

Operators also describe the (nonlinear) evolution of the coefficients $\beta(h) \mapsto \beta(ho)$:

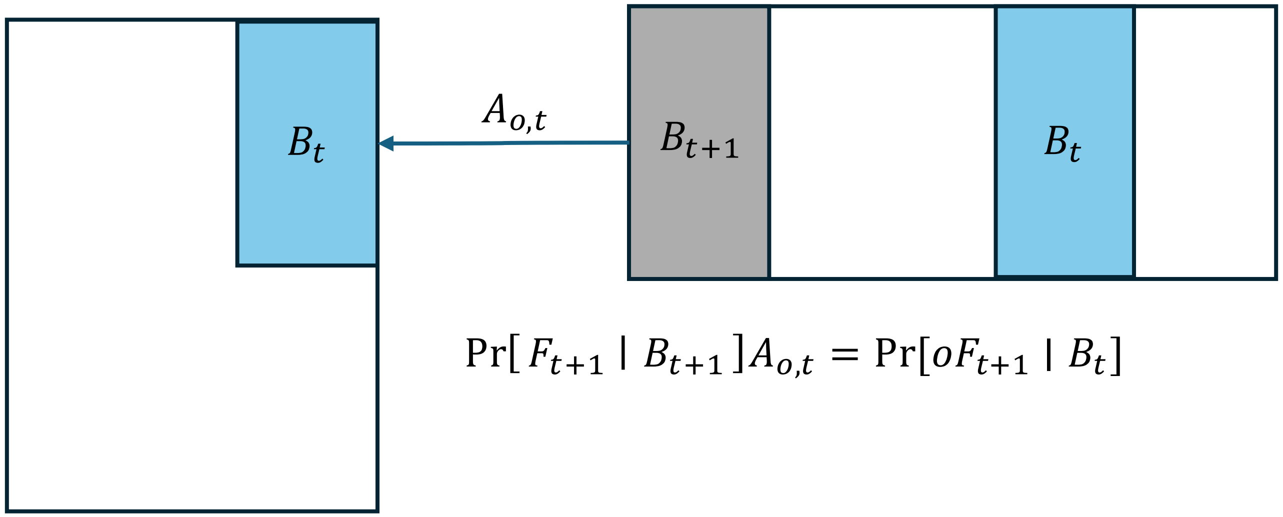
$$\Pr[o|h] \cdot \beta(ho) = A_{o,t}\beta(h)$$

Low rank distributions admit efficient/compact representation
Linear system solve yields operators when bases B_t are known (in exact model)
But how do we find the bases?

Jaeger. Observable operator models for discrete stochastic time series. 2000

Interlude: Generalizing Angluin's L*

With queries to exact cond. probs.



$$\Pr[F_{t+1} \mid B_{t+1}]A_{o,t} = \Pr[oF_{t+1} \mid B_t]$$

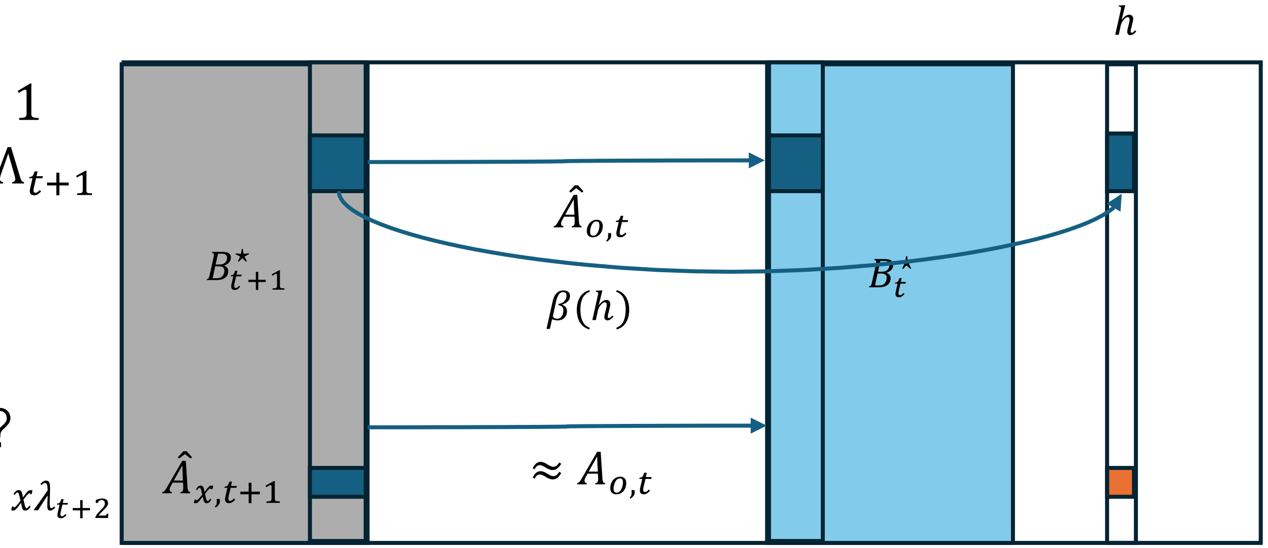
Interlude: Generalizing Angluin's L*

With queries to exact cond. probs.

Idea: grow basis iteratively, starting from $|B_{t+1}| = 1$
Replace futures F_{t+1} with spanning set of "tests" Λ_{t+1}

$$\Pr[\,\Lambda_{t+1} \mid B_{t+1}\,]\,\hat{A}_{o,t} = \Pr[o\Lambda_{t+1} \mid B_t]$$

Model is correct if rank is large, but what if it isn't?



Test: Sample sequences from $\Pr[\cdot]$, check model on $x_{1:t+1}\Lambda_{t+1}$ (for all t).

Lemma: If model agrees on $1/\epsilon^2$ sequences $x_{1:t+1}$ (for all t), we are done.

Idea: TV small if one-step errors small on average over history.

Lemma: Mistake on $hx\lambda_{t+2}$ but not $h\Lambda_{t+1}$ expands basis! Can only happen r times.

$$\Pr[\Lambda_{t+1}|h] = \Pr[\Lambda_{t+1}|B_{t+1}]\beta(h)$$
$$\Pr[x\lambda_{t+2}|B_{t+1}] = \Pr[\lambda_{t+2}|B_{t+2}]\,\hat{A}_{x,t+1}$$
$$\Pr[x\lambda_{t+2}|h] \neq \Pr[\lambda_{t+2}|B_{t+2}]\hat{A}_{x,t+1}\beta(h)$$

Proof overview

- Structural properties: compact representation via observable operators
    - Generalizing Angluin's L* algorithm with exact conditional probabilities
- Estimating operators
- Basis finding
- Error propagation

Estimating operators

Suppose we have basis histories B_t and B_{t-1}. Main equation is:

$$\Pr[F_t \mid B_t] A_{o,t-1} = \Pr[oF_t \mid B_{t-1}]$$

Issue #1: Matrices are exponentially large! Solution: use tests

Issue #2: Test probabilities $\Pr[\lambda|b]$ exponentially small; impossible to estimate with samples

Let D_t be a diagonal matrix with $d_t(f) := \mathbb{E}_{b \sim B_t}[\Pr[f \mid b]]$ on the diagonal.

$$\underbrace{\Pr[F_t \mid B_t]^\top D_t^{-1} \Pr[F_t \mid B_t]}_{\Sigma} A_{o,t-1} = \underbrace{\Pr[F_t \mid B_t]^\top D_t^{-1} \Pr[oF_t \mid B_{t-1}]}_{Q}$$

Now entries are reasonable

$$\sum_f d_t(f) \cdot \frac{\Pr[f|b_i] \Pr[f|b_j]}{d_t(f)^2}$$

Can estimate to additive accuracy, but small probabilities still tricky.

Estimating operators

$$\Sigma = \Pr[F_t \mid B_t]^\top D_t^{-1} \Pr[F_t \mid B_t]$$
$$\Sigma \cdot A_{o,t-1} = Q$$

Estimating individual entries yields Frobenius norm guarantee

$$\left|\left| \Sigma - \hat{\Sigma} \right|\right|_F \leq \gamma$$

But we need to invert $\hat{\Sigma}$ to estimate $A_{o,t-1}$. We care about singular values, could be small!

Preconditioning helps (e.g., in parity), but not always.

$$D_t^\star = \mathrm{diag}(\mathbb{E}_{b \sim B_t^\star}[\Pr[f \mid b]])$$

Fidelity: $\Pr[\cdot]$ has fidelity Δ if there exists a basis $B_t^\star$ of size $|B_t^\star| \leq 1/\Delta$ such that

$$\sigma_{\min}\left(D_t^{\star,-\frac{1}{2}} \mathbb{E}[\Pr[F_t \mid x_{1:t}] \Pr[F_t \mid x_{1:t}]^\top] D_t^{\star,-\frac{1}{2}}\right) \geq \Delta$$

Implies existence of a robust basis, one with $\sigma_{\min}(\Sigma(B)) \geq \Delta$.

Lemma: With Δ robust basis, can estimate operators $A_{o,t-1}$ in ℓ_2 norm

On fidelity

Main equation is $\Pr[\, F_t \mid B_t \,] A_{o,t-1} = \Pr[oF_t \mid B_{t-1}]$

Need to learn $A_{o,t-1}$. Essentially no other structure available!

Challenge #1: Need to estimate design matrix, already non-trivial. We use preconditioning.

Challenge #2: Small singular values => impossible to estimate $A_{o,t-1}$ in all directions.

- Can project out small directions, but unclear how these errors propagate
- High fidelity => no small singular values => can estimate $A_{o,t-1}$

Proof overview

- Structural properties: compact representation via observable operators
    - Generalizing Angluin's L* algorithm with exact conditional probabilities
- Estimating operators
- <u>Basis finding</u>
- Error propagation

Finding bases

$$\Sigma = \Pr[F_t \mid B_t]^\top D_t^{-1} \Pr[F_t \mid B_t]$$
$$\Sigma \cdot A_{o,t-1} = Q$$

Lemma: Under fidelity, a random sample of poly histories suffices

Can approximate with samples

$$\sigma_{\min}\left(D_t^{\star,-\frac{1}{2}} \mathbb{E}[\Pr[F_t \mid x_{1:t}]\Pr[F_t \mid x_{1:t}]^\top] D_t^{\star,-\frac{1}{2}}\right) \geq \Delta$$

Need to swap with $D_t(B_t)^{-\frac{1}{2}}$

Can bound $\frac{d(f)}{d^\star(f)}$ because $B^\star$ is a basis. Write $\Pr[f|h] = \Pr[f|B^\star]\alpha(h)$ with $|\alpha(h)|_2 \leq 1$

We use volumetric spanner instead of basis for norm control, one of size $O(r)$ always exists.

Can also adapt basis finding strategy from exact case, but not required under fidelity

Hazan et al. Volumetric spanners: An efficient exploration basis for learning. 2016.

Proof overview

- Structural properties: compact representation via observable operators
  - Generalizing Angluin's L* algorithm with exact conditional probabilities
- Estimating operators
- Basis finding
- Error propagation

Error propagation

$$\mathrm{TV}\big(\Pr[\cdot],\widehat{\Pr}[\cdot]\big) \coloneqq \frac{1}{2}\cdot\sum_{x_{1:T}} |\, A_{x_T} \dots A_{x_1} \; - \hat{A}_{x_T} \dots \hat{A}_{x_1} |$$

Exponentially many terms and iterated matrix multiple: error amplification!

If we have $|A_{o,t} \; - \hat{A}_{o,t}|_2 \leq O(\epsilon)$ natural to decompose

$$\mathrm{TV}\big(\Pr[\cdot],\widehat{\Pr}[\cdot]\big) \leq \sum \left|\hat{A}_{x_{T:t+2}}\right|_2 \cdot |\, \hat{A}_{x_{t+1}} - A_{x_{t+1}} \Big|_2 \left|A_{x_{t:1}}\right|_2$$

But could have $|A_{o,t}|_2 \approx r$ so terms could be exponentially large

Error propagation

$\Pr[o|h] \cdot \beta(ho) = A_{o,t}\beta(h)$

$$\mathrm{TV}\left(\Pr[\cdot], \widehat{\Pr}[\cdot]\right) := \frac{1}{2} \cdot \sum_{x_{1:T}} |\, A_{x_T} \dots A_{x_1} - \hat{A}_{x_T} \dots \hat{A}_{x_1} |$$

- Refined analysis for estimating operators $\hat{A}_{o,t}$: error in the space of coefficients

$$\left(A_{o,t} - \hat{A}_{o,t}\right)u \approx \beta(B_{t+1})\alpha \text{ (plus small orthogonal component) with } |\alpha|_1 \leq \epsilon$$

- Inductive argument with three error terms

$$A_{x_{1:t}} - \hat{A}_{x_{1:t}} = \left(A_{x_t} - \hat{A}_{x_t}\right)A_{x_{1:t-1}} + A_{x_t}\left(A_{x_{1:t-1}} - \hat{A}_{x_{1:t-1}}\right) + \left(\hat{A}_{x_t} - A_{x_t}\right)\left(A_{x_{1:t-1}} - \hat{A}_{x_{1:t-1}}\right)$$

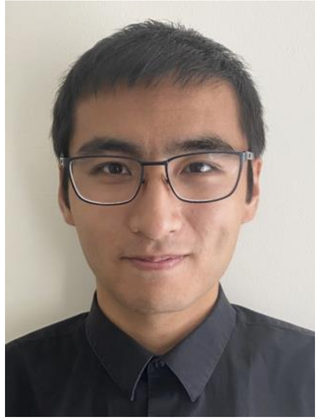
- Always track error in the space of coefficients

$$A_{x_{1:t}} - \hat{A}_{x_{1:t}} \approx \beta(H)\gamma(x_{1:t}) \text{ (plus orthogonal component) with } \textstyle\sum_{x_{1:t}} |\gamma(x_{1:t})|_1 \leq O(t\epsilon)$$

Hsu et al. A spectral algorithm for learning hidden Markov models. 2009.

Conclusion and discussion

- Recap: Interactive access (cond. probs. or samples) can bypass computational hardness for HMMs
  - All HMMs with conditional probability access
  - HMMs with high fidelity with conditional samples: covers parity but not all DFAs
- Open problem: Efficiently learn all HMMs with conditional samples
  - Challenge is poor conditioning: cannot estimate operators $A_{o,t}$ in all directions
  - But truncation/projection poorly understood: approximate an HMM by one with fewer states?
- Practical speculation: Can conditional sampling improve LLMs?



Thanks!